Event Information Extraction

Extracting role fillers associated with events.

Examples

Terrorism: perpetrator, victim, target, date, location

Management succession: person fired, successor, position, organization, date

Disease outbreaks: disease, victim, symptoms, containment measures

Sample Text

Event type: bombing

Perpetrator: guerrilla urban commandos

Weapon: highpower bombs

Target: car dealership

Location: San Salvador

Date: this morning

Damage: set the building on fire, but did not result any injury

Filled Terrorism Event Template

<table>
<thead>
<tr>
<th>Date</th>
<th>10 January 1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>El Salvador: San Salvador (city)</td>
</tr>
<tr>
<td>Event type</td>
<td>bombing</td>
</tr>
<tr>
<td>Weapon</td>
<td>highpower bombs</td>
</tr>
<tr>
<td>Perpetrator individual</td>
<td>guerrilla urban commandos</td>
</tr>
<tr>
<td>Perpetrator organization</td>
<td></td>
</tr>
<tr>
<td>Physical target</td>
<td>car dealership</td>
</tr>
<tr>
<td>Physical target effect</td>
<td>some damage</td>
</tr>
<tr>
<td>Human target</td>
<td></td>
</tr>
<tr>
<td>Human target effect</td>
<td>no injury or death</td>
</tr>
</tbody>
</table>

IE in the Wild: ProMed Disease Outbreak Reports

Elapsed time: 00:23:28
Ebola Haemorrhagic Fever In Uganda - Update 5

As of Sat 21 Oct 2000, the Ugandan Ministry of Health has reported 139 cases including 51 deaths. The increase of 17 cases in the last 24 hours reflects the intensified active surveillance.

A team from the WHO Collaborating Centre at the US Centers for Disease Control and Prevention (CDC), United States is establishing a field diagnostic laboratory in Gulu district. The last laboratory equipment arrived Sat 20 Oct 2000 and the laboratory is expected to be operational shortly. A WHO information officer from Geneva arrived in Uganda on Wed 18 Oct 2000 and is based in Gulu district. He is working with the Ugandan Ministry of Health as media focal point.

Ebola Haemorrhagic Fever In Uganda - Update 6

As of Sun 22 Oct 2000, the Ugandan Ministry of Health has reported 149 cases, including 54 deaths. [This represents an increase of 10 cases and 3 deaths in the last 24 hours. - Mod_CP]

Disease Outbreak Template Definition

<table>
<thead>
<tr>
<th>Story:</th>
<th>&lt;document id&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID:</td>
<td>&lt;template id&gt;</td>
</tr>
<tr>
<td>Date:</td>
<td>&lt;date&gt;</td>
</tr>
<tr>
<td>Event:</td>
<td>outbreak</td>
</tr>
<tr>
<td>Status:</td>
<td>confirmed, suspected, or possible</td>
</tr>
<tr>
<td>Containment:</td>
<td>culling, disinfecting, facility closing, inspection, medicine, pesticide, quarantine, vaccine, or other</td>
</tr>
<tr>
<td>Country:</td>
<td>&lt;set fill&gt;</td>
</tr>
<tr>
<td>Victims:</td>
<td>&lt;string list&gt;</td>
</tr>
<tr>
<td>Disease:</td>
<td>&lt;string&gt;</td>
</tr>
</tbody>
</table>

Filled Disease Outbreak Template

<table>
<thead>
<tr>
<th>Story:</th>
<th>20020714.4756</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID:</td>
<td>1</td>
</tr>
<tr>
<td>Date:</td>
<td>August 14, 2002</td>
</tr>
<tr>
<td>Event:</td>
<td>outbreak</td>
</tr>
<tr>
<td>Status:</td>
<td>confirmed</td>
</tr>
<tr>
<td>Containment:</td>
<td>none</td>
</tr>
<tr>
<td>Country:</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Victims:</td>
<td>the 27 reported cases</td>
</tr>
<tr>
<td>Disease:</td>
<td>Creutzfeldt-Jakob Disease / [sporadic] Creutzfeldt-Jakob disease (CJD) / CJD / Sporadic CJD / hereditary dominant CJD / Swiss CJD / sporadic Creutzfeldt-Jakob disease</td>
</tr>
</tbody>
</table>
Event Extraction vs. Named Entity Recognition

- **Named Entity Recognition** = identifying types of entities
- **Event Extraction** = identifying role relationships associated with events.

Paul Nelson killed John Smith.
Paul Nelson was killed by John Smith.

IBM purchased Microsoft.
IBM was purchased by Microsoft.
IBM was purchased on Tuesday by Microsoft.

Unstructured vs. Semi-structured Text

Unstructured text depends 100% on language understanding. Semi-structured text has some structure (layout) that can aid in understanding.

**Unstructured Text**
Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled “Embedded Commitment,” on Thursday, May 4th from 4-5:30 in PH 223D.

**Semi-Structured Text**
Laura Petitte
Department of Psychology
McGill University
Thursday, May 4, 1995
12:00 pm
Baker Hall 355

Another Semi-Structured Seminar Announcement

Name: Dr. Jeffrey D. Hermes
Affiliation: Department of AutoImmune Diseases Research & Biophysical Chemistry Merch Research Laboratories
Title: “MHC Class II: A Target for Specific Immunomodulation of the Immune Response”
Host/e-mail: Robert Murphy
Date: Wednesday, May 3, 1995
Time: 3:30 p.m.
Place: Mellon Institute Conference Room
Sponsor: MERCK RESEARCH LABORATORIES

Patterns/Rules vs. Sequence Tagging

Two general approaches to IE:

*Pattern-based systems* use patterns or rules that are applied to text.

*Sequence tagging models* classify individual tokens as to whether or not they should be extracted.
IBM fired its CEO.

**Event:** FIRING  
**Agent:** IBM  
**Fired:** its CEO

John Smith was let go on Monday.

**Event:** FIRING  
**Agent:** IBM  
**Fired:** John Smith, CEO  
**Date:** Monday

---

### Pattern-based Template-Filling Pipeline

1. **Syntactic Analysis**
   - IBM fired its CEO.

2. **Pattern Extraction**
   - Event: FIRING  
   - Agent: IBM  
   - Fired: its CEO

3. **Coreference**
   - IBM fired its CEO.  
   - John Smith was let go on Monday.

4. **Template Creation**

---

### Case Frames

5000 people were infected with the Ebola virus in a horrific outbreak in western Africa.

**Trigger:** (passive-vp “infected”)  
**Event:** OUTBREAK  

<table>
<thead>
<tr>
<th>SynRole</th>
<th>EventRole</th>
<th>Semantic</th>
<th>Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject</td>
<td>victim</td>
<td>animate</td>
<td>5000 people</td>
</tr>
<tr>
<td>PP(with)</td>
<td>disease</td>
<td>disease</td>
<td>the Ebola virus</td>
</tr>
<tr>
<td>PP(by)</td>
<td>disease</td>
<td>disease</td>
<td></td>
</tr>
<tr>
<td>PP(in)</td>
<td>location</td>
<td>place</td>
<td>western Africa</td>
</tr>
<tr>
<td>PP(in)</td>
<td>date</td>
<td>time</td>
<td></td>
</tr>
</tbody>
</table>

---

### Lexico-Syntactic Patterns

Five priests were murdered by armed guerrillas.

**Trigger:** VP-passive(“murdered”)  
**EventType:** MURDER  
**Victim:** Subject = Five priests  
**Semantic Constraints:** {HUMAN}

**Perpetrator:** PP(by) = armed guerrillas  
**Semantic Constraints:** {HUMAN}

---

### Supervised Learning for IE

- In the mid-1990s, researchers began to develop methods to automatically create (learn) IE systems.
- Supervised learning requires annotated training data.
- **Trade-off:** annotating texts vs. manual knowledge engineering  
  - weeks vs. months  
  - domain experts vs. computational linguists
Alleged guerrilla urban commandos launched two highpower bombs against a car dealership in downtown San Salvador this morning. A police report said that the attack set the building on fire, but did not result any casualties.

IE as Sequence Tagging

- A different approach: build a classifier as a sequence tagging model.
- Each document is processed sequentially and each token is labeled as Extraction or Non-Extraction.
  Ex: B (beginning), I (inside), or O (outside) tags.
- Features are usually simple: e.g., words, POS tags, orthography, and a small context window of preceding/following words.

Sequence Tagging Example

Alleged guerrilla urban commandos launched two highpower bombs against a car dealership in downtown San Salvador this morning. A police report said that the attack set the building on fire, but did not result any casualties.

AutoSlog [Riloff 1993]
The Perils of Manual Text Annotation

- Time consuming
- Tedious
- Deceptively tricky
- A new corpus must be annotated for each domain!

Weakly Supervised Learning for IE

- Idea: can we train an IE system using only unannotated texts?

  - Yes, if we have “preclassified” texts:
    - One pile of relevant texts
    - One pile of irrelevant texts
    - Manual review of ranked patterns

  - Much easier than annotating texts!

AutoSlog-TS [Riloff 96]
(Step 1)

Relevant

[The World Trade Center], [an icon] of [New York City], was horrifically attacked on [an otherwise beautiful day] in [September 2001] by [Al Qaeda].

Irrelevant

Shallow Parser

Syntactic Templates

Extraction Patterns:
- <subj> was attacked
- icon of <np>
- was attacked on <np>
- was attacked in <np>
- was attacked by <np>
AutoSlog-TS (Step 2)

Relevant

Extraction Patterns:

- <subj> was attacked
- icon of <np>
- was attacked on <np>
- was attacked in <np>
- was attacked by <np>

<table>
<thead>
<tr>
<th>Extraction Patterns</th>
<th>Freq</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;subj&gt; was attacked</td>
<td>100</td>
<td>.90</td>
</tr>
<tr>
<td>icon of &lt;np&gt;</td>
<td>5</td>
<td>.20</td>
</tr>
<tr>
<td>was attacked on &lt;np&gt;</td>
<td>80</td>
<td>.79</td>
</tr>
<tr>
<td>was attacked in &lt;np&gt;</td>
<td>85</td>
<td>.87</td>
</tr>
<tr>
<td>was attacked by &lt;np&gt;</td>
<td>95</td>
<td>.95</td>
</tr>
</tbody>
</table>

Irrelevant

Examples of Learned Disease Patterns

- outbreak of <np>
- <subj> spread
- cases of <np>
- <subj> was confirmed
- outbreaks of <np>

- <subj> was transmitted
- contracted <dobj>
- spread of <np>
- <subj> infected
- <subj> killed

Top Terrorism Extraction Patterns

1. <subject> exploded
2. murder of <np>
3. assassination of <np>
4. <subject> was killed
5. <subject> was kidnapped
6. attack on <np>
7. <subject> was injured
8. exploded in <np>
9. death of <np>
10. <subject> took_place
11. caused <dobj>
12. claimed <dobj>
13. <subject> was wounded
14. <subject> occurred
15. <subject> was located
16. took_place on <np>
17. responsibility for <np>
18. occurred on <np>
19. was wounded in <np>
20. destroyed <dobj>
21. <subject> was murdered
22. one of <np>
23. <subject> kidnapped
24. exploded on <np>
25. <subject> died

The Bootstrapping Era

Unannotated Texts

Unannotated Texts + = KNOWLEDGE!
Learning Extraction Patterns

Seed Patterns

Unannotated

AutoSlog-TS

Relevant Irrelevant

Syntactic Rules

Manual Review

best pattern ranked patterns final patterns

ExDisco [Yangarber et al. 2000]

Secondary Contexts

A terrorist arrested by the Salvadoran national police, has been identified as Ruth Esperanza Aguilar Marroquin.

Oqueli’s body was found next to the body of Guatemalan politician Gilda Flores.

There were seven children, including four of the Vice President’s children, in the home at the time.

Event Keywords

Keywords alone are not as reliable as you might think due to ambiguity, metaphor, and context.

The comedian bombed at the club …

Parliament exploded in anger about …

Obama was attacked by House Republicans …

Linker [Huang & Riloff 2012]
### Event Extraction Performance

<table>
<thead>
<tr>
<th>System</th>
<th>Average (P/R/F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLACIER (2009)</td>
<td>48/57/52</td>
</tr>
<tr>
<td>TIER (2011)</td>
<td>51/62/56</td>
</tr>
<tr>
<td>Linker (2012)</td>
<td>58/60/59</td>
</tr>
</tbody>
</table>

### Challenges for the Future

- **Contextual Effects**

  *The man took the money and fled.*

  - robbery

  *Abilio Diniz is in the hands of a group presumed to be Chilean terrorists.*

  - kidnapping

- **Inference**

  *He was shot.*
  *He was shot to death.*
  *He was riddled with machine gun fire.*
  *He was shot. His body was found yesterday.*

- **Metaphor**

  *killing two birds with one stone*
  *shot in the arm; shot in the dark*